Method of Mixed Traffic Model Formation

Alexey Begaev
North West Echelon, JSC
St-Petersburg, Russia
a.begaev@nwechelon.ru

Mikhail Chesnakov, Yuriy Starodubtsev
32nd Department
Budyonny Military Academy of Communications,
St-Petersburg, Russia,
chesnakof@gmail.com; ys@e-nw.ru

***Abstract* — This paper proposes a method of mixed traffic model formation, which allows to create statistical models of mixed traffic for each network element as well as to have sustainable statistical data characterizing mixed traffic on each network element. It also shows diversity of network traffic by its basic characteristics. Limitations of applicability of existing models and methods complex for mixed network traffic specification are discussed herein. We offer a variant of mixed stream decomposition to uniform streams using the Theory of Pattern Recognition methods. We offer a variant of uniform stream representation as random numerical sequences relevant to packets arrival time. There is grounding for selection of rules for checking of accordance between experimental and theoretical distribution in respect to uniform streams of network traffic typical for existing and perspective information telecommunication systems.**

Keywords — traffic; stream; model; network; information telecommunication systems; Theory of Pattern Recognition; statistical analysis; random values distribution law

I. Introduction

The rationale for developing of method of mixed traffic model formation is predetermined by significant number of actual circumstances and importance of date characterizing traffic for practice.

These data are necessary for solving important practical tasks on calculation of probabilistic time-response characteristics of specified subnetwork elements, required performance determination – μ, at specified traffic intensity – λ and at assigned service procedure by switching nodes, and finding facts and reasons for traffic parameters abnormal alteration [1].

Relating to the statistical and uniform traffic, we developed a complex of models and methods [2, 3, 4, 5] which allows to solve practical tasks with adequate accuracy.

However, the current multiservice communication systems are characterized by a number of distinguishing features which do not admit of traditional methodological approach.

Actual information telecommunication systems are built and operated by a significant number of operators using hardware and software from various manufacturers. The situation is characterized by continuous development of technical specifications and standards used in information telecommunication systems while manufactures implement their different versions [6].

Various routing options as well as destruction actions of individual intruders (hackers) and their organized groups have a great impact on traffic parameters [12].

Consequently, current information telecommunication systems traffic is mixed and highly dynamic. Herewith it may dramatically differ at various network points [7].

The method allows to have stable statistical date characterizing mixed traffic for each element of specified subnetwork element of communication network.

Standard approaches based on mathematical statistics methods cannot be applied because they do not provide event stream uniformity. Network packets differ from each other on a variety of characteristics: type, size, address, priority, etc. Request for communication service processing includes connection request stream as well as stream of transmitted user's information.

Further on, it is expected that in properly functioning network the time share for connection request processing in the overall traffic volume is substantially less than time share for data exchange. All the switching nodes have the same service procedure.

Clear representation about the scope of information processed by switching nodes in current information telecommunication systems can be obtained examining statistic of overall traffic transferred through Internet Exchange Points. The overall traffic transferred through node MSK-IX[1] is shown on Figure 1 [8].

[1] https://www.msk-ix.ru/traffic/

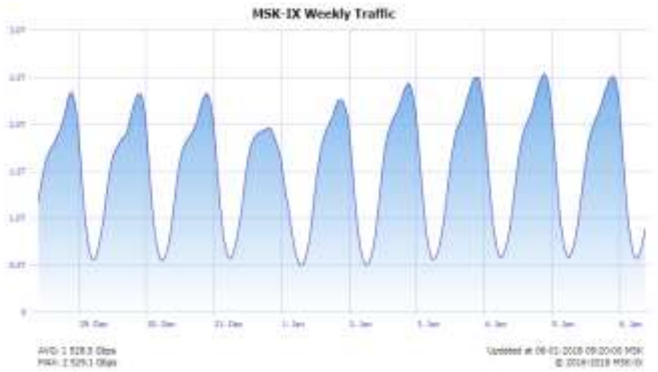

Fig. 1. Member’s overall traffic transferred through node MSK-IX.

II. Method of Mixed Traffic Model Formation

Significant traffic volumes allow to have a huge sampling that is substantially different from the situation where the number of experiments is relatively small.

The method of mixed traffic model formation is expected to be realized in relatively stand-alone five phases. Graphical view of model formation process is shown on Figure 2.

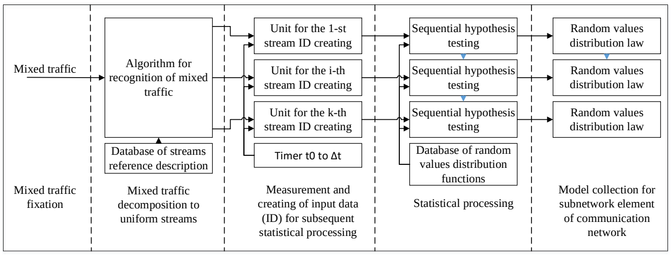
Mixed traffic
Algorithm for recognition of mixed traffic
Database of streams reference description
Unit for the 1-st stream ID creating
Unit for the i-th stream ID creating
Unit for the k-th stream ID creating
Timer t0 to Δt
Sequential hypothesis testing
Sequential hypothesis testing
Sequential hypothesis testing
Database of random values distribution functions
Random values distribution law
Random values distribution law
Random values distribution law
Mixed traffic fixation
Mixed traffic decomposition to uniform streams
Measurement and creating of input data (ID) for subsequent statistical processing
Statistical processing
Model collection for subnetwork element of communication network

Fig. 2. Mixed traffic modeling process.

The first phase involves elements fixation for traffic processed by *i*–th element of selected communication subnetwork.

There are dedicated means — network protocol analyzers which are used for mixed traffic fixation and resulted IP packets header values determination. The typical functions of network protocol analyzer are packets capturing, decryption, packet analysis and displaying. As an example the most common network protocol analyzers could be considered: Wireshark, York, SoftPerfect Network Protocol Analyzer, Accurate Network Monitor and etc. All of them allow to have information concerning date and time of packet capturing, source and destination IP address, protocol type (network, transport or application layer) and other information about captured data.

During the second phase, based on the set of specified characteristics a mixed traffic stream is decomposed to uniform ones. The model is suitable for various network protocols traffic processing, at the same time packet header formats may differ by structure. The header formats have considerable number of fields which can take on considerable but limited number of values. Decomposition based on packet classification condition with exactly identical values in all fields will lead to unnecessarily increasing of uniform stream number that make more difficult to realize proposed model.

Based on existing tasks it is acceptable to ignore some of fields values. From the other hand it is possible to perform packets classification conditions according to all header fields values. Moreover we can use specific traffic analyzers to make classification according to packet body content. It confirms the model flexibility.

We present the mixed traffic stream in a form of some data aggregate. In the proposed model decomposition is based only on header characteristics, characteristics of payload transferred in packet when assigning stream to certain class will be ignored.

The recognition performs two basic operations. At first, it is calculation of realization similarity factor with all references. Second operation is assigning of realization to reference with highest similarity. The recognition as decomposition of some set to certain number of non-empty disjoint subsets using selected criteria.

The primary criterion is the assignment of mixed traffic stream to one of existing network protocol (IP, X.25, etc.), at a later stage classification is performed based on criteria arising from differences of packet header fields values. Network packet structure and fill range of permissible fields values are always known and finite which is necessary condition for combining of various networks to single one. Up to date described in RFC 791 specification IPv4 protocol and its sequel, IPv6, are basic network protocols. This protocol is used as an example for further description but the developed method allows to work with any primary date.

IP packet header size may vary from 20 bytes to 60 bytes and contain as minimum 12 fields (Version, IHL, Type of Service, Total Length, Identification, Flags, Fragment Offset, Time to Live, Protocol, Header Checksum, Source Address, Destination Address), therefore, assignment of packets with identical headers to separate class would create its huge number.

Format of IP packet header is shown on Figure 3.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version|  IHL  |Type of Service|          Total Length         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Identification        |Flags|      Fragment Offset    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Time to Live |    Protocol   |         Header Checksum       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Source Address                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Destination Address                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Options                    |    Padding    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Fig. 3. Format of IP packet header.

IP packet header fields description:

- Version: 4 bits. The Version field indicates the format of the internet header.
- IHL: 4 bits. Internet Header Length is the length of the internet header in 32 bit words, and thus points to the beginning of the data.
- Type of Service: 8 bits. The Type of Service provides an indication of the abstract parameters of the quality of service desired.
- Total Length: 16 bits. Total Length is the length of the datagram, measured in octets, including internet header and data. This field allows the length of a datagram to be up to 65,535 octets.
- Identification: 16 bits. An identifying value assigned by the sender to aid in assembling the fragments of a datagram.
- Flags: 3 bits. Various Control Flags.
- Time to Live: 8 bits. This field indicates the maximum time the datagram is allowed to remain in the internet system. If this field contains the value zero, then the datagram must be destroyed. This field is modified in internet header processing.
- Protocol: 8 bits. This field indicates the next level protocol used in the data portion of the internet datagram.
- Options: variable. The options may appear or not in datagrams.

Full description of IP packet header fields you can find at RFC 791[2].

In the context of the current task the subject of interest is only uniform streams with a large share in overall stream. It is reasonable to group all relatively uncommon streams into separate class.

Packet reference description database may be presented in logic table format. Let's define by I a set containing selected classes of homogeneous in the sense of equality of header selected fields values or disjoint values ranges, and by J a set of all possible header fields values or disjoint values ranges. In this case if j-th header field value corresponds to i-th class of packets then table element $k_{IJ}(i,j) = 1$, otherwise $k_{Ij}(i,j) = 0$.

A table such as the one described above but containing all possible variants of values would have dramatic dimension that is not necessary because in practice only packet classes containing certain values in header are interesting.

To assign any mixed traffic packet to closest uniform class we will use the Theory of Pattern Recognition methods. The Theory of Pattern Recognition method based on pair-wise comparison of object to be recognized with reference set. The following similarity measures are available for binary data[3]:

Russell-Rao. This is a binary version of the inner (dot) product. Equal weight is given to matches and nonmatches. This is the default for binary similarity data.

Simple matching. This is the ratio of matches to the total number of values. Equal weight is given to matches and nonmatches.

Jaccard. This is an index in which joint absences are excluded from consideration. Equal weight is given to matches and nonmatches. Also known as the similarity ratio.

Dice. This is an index in which joint absences are excluded from consideration, and matches are weighted double. Also known as the Czekanowski or Sorensen measure.

Rogers and Tanimoto. This is an index in which double weight is given to nonmatches and others.

The indices listed above can be used as a function $J(Y_1, Y_2, \dots, Y_q)$, which determines the "distance" between classes in the attribute space with the coordinates $Y_1, Y_2, \dots, Y_q$.

The task of pattern recognition using the methods of statistical recognition theory is realized in two stages. The stage of learning and constructing the standard descriptions of classes and the stage of recognition.

The source of information about recognizable images is the set of results of independent observations (sampling values) that make up the learning (learning) $(x_i)1^{m_k} = (x_1, x_2, \dots, x_{m_k})$ and the control (exam) $(x_i)1^n = (x_1, x_2, \dots, x_n)$ samples, and depending on the nature of the recognition problem (one-dimensional or multidimensional) x_i can be either a one-dimensional or a p- dimensional random variable.

Training is aimed at the formation of standard class descriptions. The decisive rule based on the formation of the likelihood ratio and its comparison with a certain threshold c, the value of which is determined by the selected quality criterion:

$$\hat{L} = \frac{\widehat{\omega}_n(x_1, x_2, \dots, x_n | s_2)}{\widehat{\omega}_n(x_1, x_2, \dots, x_n | s_1)} \geq c \qquad (1)$$

where $\widehat{\omega}_n(x_1, x_2, \dots, x_n | s_j)$ is the he estimate of the conditional joint n-dimensional probability density $x_1, x_2, \dots, x_n$ provided they belong to the class s_j.

At the stage of training and the construction of reference class descriptions, the following actions are performed:

1) Form a set of characteristics from the number of available to measure the characteristics of the object $Y_1, Y_2, \dots, Y_q$.
2) Specify the function $J(Y_1, Y_2, \dots, Y_q)$ that defines the "distance" between classes in the characteristic space with the coordinates $Y_1, Y_2, \dots, Y_q$.

[2] https://tools.ietf.org/html/rfc791

[3] www.ibm.com/support/knowledgecenter/ru/SSLVMB_24.0.0/spss/base/cmd_proximities_sim_measure_binary.html

3) Define the probability distribution of probability characteristics for classes.

4) Calculate and select p new characteristics $X_1, X_2, \dots, X_p$, $p < q$, which correspond to the minimal eigenvalues λ_j in the sum $J = 2tr|M = \sum_{j=1}^{p} \lambda_j,$).

The above sequence of actions will reduce the number of features that will reduce the cost of performing measurements and calculations.

The recognition problem can be reduced to the problem of recognition of multidimensional normal populations. Approaches to the solution of this problem are clearly set forth in [9].

At the stage of measuring and creating of primary data for further statistical processing the random numerical sequences relevant to arrival time of packets belonging to uniform stream will be received in a form of sequences of arrival times of packets belonging to uniform streams: $T^i(t_s; t_s + \Delta t) = \{T_1^i, \dots, T_{\mathrm{l}}^i, \dots, T_p^i\}$,, where T^i - numerical sequence of arrival times of packets belonging to uniform stream; $t_s; t_s + \Delta t$ - current time range; T_1^i - arrival time of l-th packet, i-th stream.

The selection of set of distribution functions was conducted on the basis of physical meaning of random value specifying time intervals between uniform traffic packets arrivals. Random values will be located only on positive semiaxis and uniform by nature traffic for which IP-header fields values are equal may be overall traffic of large number of users or applications used one type communication service.

Database of distribution functions may be created from following distribution laws: gamma distribution, Erlang distribution, Rayleigh distribution, Pareto distribution and others which are not contrary to physical meaning of random value specifying time intervals between uniform traffic packets arrivals.

Mentioned above distribution laws are presented in Table 1.

TABLE 1. DENSITY DISTRIBUTION FUNCTIONS

Distribution function name	Density Distribution
Gamma distribution	$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x > 0,$ where λ – scale parameter (λ>0); α – shape parameter (α>0)
Erlang distribution of m-th order	$f(x) = \frac{\lambda^m}{(m-1)!} x^{m-1} e^{-\lambda x}, x \geq 0,$ where λ – scale parameter (λ>0); m – shape parameter, distribution order, positive real number (m≥ 1)
Rayleigh	$f(x) = \frac{x}{a^2} e^{-x^2/(2a^2)}, x > 0,$ where a – scale parameter, mode (a>0)
Pareto	$f(x) = \frac{\alpha}{x_0} \left(\frac{x_0}{x}\right)^{\alpha+1}, x > x_0,$ where x_0 – location parameter, left border of possible values range ($x_0 > 0$); α – shape parameter (α>0)

During fourth stage statistical processing of uniform network traffic streams performs to establish continuous distribution law which most highly specifies random value sample of which was obtained during experimental observations, a hypothesize concerning accordance between experimental and theoretical distribution put forward which may be checked applying various accordance criteria [9].

The most frequently applicable in practice criteria are: 1) Criteria of χ^2 type; 2) Various non-parametric criteria: Kolmogorov criterion, Smirnov criterion, Mises criterion. They differed in the conditions of applicability when testing the accordance hypothesize for various distribution laws (see GOST R 50.1).

There is difference between simple and complex hypothesizes. The simple tested hypothesize has a form: $H_0: f(x) = f(x, \theta_0)$, where $f(x)$ - density function; θ_0 - known scalar or vector parameter of theoretical distribution which used during accordance testing. The complex hypothesize has a form $H_0: f(x) \in \{f(x, \theta), \theta \in \Theta\}$, where Θ – space of parameters and scalar or vector parameter estimator $\hat{\theta}$ is calculated using the same sampling as for accordance hypothesize testing [11, 12].

From the proposed in the method sequence of events taking into account characteristics of obtained uniform traffic streams and applicability of various accordance criteria we offer use hypothesize testing criterion of χ^2 type for testing accordance between experimental and theoretical distribution. Application of χ^2 type criteria is described in GOST R 50.1.

When testing simple hypothesize concerning accordance between experimental and theoretical distribution of random value X, the following sequence of actions is implemented:

a) Form a tested hypothesize by choosing a theoretical distribution of random value $F(x, \theta)$ accordance of which is worth checking.

b) Make random sampling of N volume from aggregation.

c) According to sampling volume N select interval number k.

d) Select edge points of group interval. In doing so the sampling may be stratified into intervals of equal length, intervals of equal probability or according to asymptotically optimum grouping for selected distribution law, but because distribution laws for various Δt may be different, we suggest to use the stratifying into intervals of equal length. In this case it is necessary to calculate number n_i and determine probability values $P_i(\theta)$.

e) After calculations n_i and $P_i(\theta)$ according to selected testing criterion it is necessary to calculate test statistics value S^* according to the formula (2) or (3):

$$S_{\chi^2} = N \sum_{i=1}^{k} \frac{(n_i/N - P_i(\theta))^2}{P_i(\theta)}, \quad (2)$$

$$S_{\text{оп}} = -2 \ln l = -2 \sum_{i=1}^{k} n_i \ln\left(\frac{P_i(\theta)}{n_i/N}\right). \quad (3)$$

f) According to χ_{k-1}^2 - distribution in accordance with the formula (4) calculate value $P\{S > S^*\}$. If $P\{S > S^*\} > \alpha$, where α is specified significance level, then there is no reason for rejecting of tested hypothesize. Otherwise, tested hypothesize is rejected.

$$P\left\{S_{\chi^2} > S^*_{\chi^2}\right\} = \frac{1}{2^{r/2}\Gamma(r/2)}\int_{S_{\chi^2}}^{\infty} S^{r/2-1}\, e^{-s/2} ds > \alpha \quad (4).$$

Calculated test statistics value S^* is compared with critical value $S_{r,\alpha}$, where $r = k - 1$ is the number of degrees of freedom defined by the equation:

$$\frac{1}{2^{r/2}\Gamma(r/2)}\int_{S_{r,\alpha}}^{\infty} S^{r/2-1} e^{-s/2} ds = \alpha. \quad (5)$$

Values $S_{r,\alpha}$ are given in the various handbooks. Accordance hypothesize is rejected if test statistics value is in critical range, i.e. at $S^* > S_{r,\alpha}$.

During complex hypothesize testing and parameter estimators calculation on grouped date, as a result of minimization of statistics predetermined by formulas (2) and (3) a checking sequence is similar to case of simple hypothesize with setting the number of degrees of freedom $r = k - 1$, where m is number of parameters estimated according to this sampling. Herewith, recommendations regarding grouping method remain valid.

At the firth stage reasonable set of distribution functions is received, each function specify particular network traffic packet stream as well as their aggregate traffic source.

III. Conclusions

Developed method allows to:

- Create statistical models of mixed traffic for each network element.
- Obtain statistical models of mixed traffic which can be used for analysis of real communication networks and design of perspective communication networks.
- Provide its updating when implementing perspective protocols.

With additional development of methods of model inter-comparison for various network elements obtained using proposed method fix the fact of abnormal traffic change and identify its reasons.

References

[1] Staroduvtsev Yu.I., Begaev A.N., Davlyatova M.A. Quality Management of Information Services. – SPb: SPbSTU, 2017, 454p. (In Russ.).

[2] Anisimov V.V., Begaev A.N., Staroduvtsev Yu.I. Functional model of communication network with unknown level of confidence and assess its capabilities to provide VPN service with specified quality. Voprosy kiberbezopasnosti *[Cybersecurity issues]*. 2017. N 1 (19), pp. 6-15. DOI: 10.21681/2311-3456-2017-1-6-15.

[3] Gross D.,Shortle J.F., Thompson J.M., Harris C.M. Fundamentals of Queueing Theory. 4th Ed. Wiley-Interscience, 2008, 528 p.

[4] Krylov V.V., Samohvalov S.S. Teletraffic and its application theory. – Spt.: BHV - Peterburg, 2005 – 288 p. (In Russ.).

[5] Starodubtsev Yu.I., Begaev A.N., Kozachok A.V. The method of controlling access to information resources of multi-service networks of various levels of confidentiality. Voprosy kiberbezopasnosti *[Cybersecurity issues]*. 2016. N 3 (16), pp. 13-17.

[6] Markov A., Luchin D., Rautkin Y., Tsirlov V. Evolution of a Radio Telecommunication Hardware-Software Certification Paradigm in Accordance with Information Security Requirements. In Proceedings of the 11th International Siberian Conference on Control and Communications (Omsk, Russia, May 21-23, 2015). SIBCON-2015. IEEE, 2015, pp. 1-4. DOI: 10.1109/SIBCON.2015.7147139.

[7] Vencel E.S.The theory of probability: Textbook for university students. 9-th ster. ed. - M.: Publishing House "Academia", 2003. - 576 p. (In Russ.).

[8] Buranova M.A. Analysis of statistical characteristics of multimedia traffic aggregation node in a multiservice network. / M.A. Buranova, V.G. Kartashevsky, M.S. Samoilov. // Radio-technical and telecommunication systems. Systems, networks and devices of telecommunications. -Murom, 2014. - No 4 (16). - P. 63-69. (In Russ.).

[9] Y.A. Fomin, G.R. Tarlovskii. Statistical Theory of Recognition of Images. - M .: Radio and Communication, 1986. -264 p. (In Russ.).

[10] Anisimov V.V., Begaev A.N., Starodubtsev Yu.I., Sukhorukova E.V., Fedorov V.G., Chukarikov A.G., The way of purposeful transformation of the model parameters of the real fragment of the communication network. Printed: May 23, 2016, Bul. N 15, 2620200. (In Russ.).

[11] Begaev A.N., Starodubtsev Yu.I., Fedorov V.G.. A method for estimating the manageability of a fragment of a public communication network, taking into account the influence of a plurality of control centers and destructive program influences. Voprosy kiberbezopasnosti *[Cybersecurity issues]*. 2017 N 4 (22), pp. 32-39. DOI: 10.21681/2311-3456-2017-4-32-39.

[12] Starodubtsev Yu.I., Grechishnikov E.V., Komolov D.V. Use of neural networks to ensure stability of communication networks in conditions of external impacts. Telecommunications and Radio Engineering. 2011. V. 70. N 14. P. 1263-1275.